Reply: Unexpected Duplicate Ship Reports in the Comprehensive Ocean-Atmosphere Data Set (COADS)

Background

Although the elimination of duplicate ship observations appears to be a very simple task, it is actually one of the more difficult aspects of surface marine data-processing. Ship reports are received from many data sources and through a variety of methods. Different conversion techniques, keypunching errors, or different interpretations among maritime nations occasionally cause duplicate ship reports which come from more than one source (but from the same logbook) to be slightly different. The advent of telecommunications data in the 1960s enlarged this problem. Reports transmitted directly from a ship would often have a high percentage of keying errors or transmission problems which alter observations.

The duplicate elimination (DUPELIM) procedure developed for the Comprehensive Ocean-Atmosphere Data Set (COADS) is very complex and is designed according to the specific data problems which emerged through two years of extensive testing (Slutz *et al.*, 1985, hereafter referred to as *Release 1)*. Based upon these tests, it was discovered that all duplicates could not be eliminated from marine data using current technology. Random errors due to keypunching could place an observation 100 years or a thousand miles away from a correctly keyed duplicate. Fortunately, most errors in marine data are more systematic and can be eliminated by special DUPELIM allowances.

Potential duplicates in COADS

Lander and Morrissey (1987, hereafter referred to as "the authors") sampled approximately 98,000 ship reports in the equatorial western Pacific. The data were for the decade 1970-79 and came from COADS Compressed Marine Reports (CMR; see Supplement D in *Release 1).* Of the ship reports sampled, 603 pairs of potential duplicates were found (a duplication rate of about 0.6%). The authors were kind enough to send a complete print-out of all pairs of potential duplicates for our inspection. These data were examined with the following results:

(1) An error existed in the authors' processing scheme which located the duplicates (Lander and Morrissey, personal communication). This error had the effect of overestimating the number of duplicates by approximately 7%.

(2) Of the remaining duplicate pairs not included in item (1), most involved telecommunications (telecom) source decks (*i.e.*, at least one record of each duplicate pair was usually a telecom report). These were predominantly deck 555 (Monterey) and deck 888 (U.S. Air Force Global Weather Center or GWC). Most duplicates were exact matches except for usually whole degree differences in latitude and/or longitude.

(3) Of the duplicates in item (2), it was anticipated that most should have been flagged as suspect during the National Climatic Data Center (NCDC) quality control (QC) track check operations which were performed on some 1970-79 data in COADS. To verify the assumption made in item (3), a random sample of 10 of the authors' duplicate pairs were selected. Since QC flag fields are not present in CMR (although they do contain "trimming flags"), another similar COADS product, which contains QC flags, was used: NCDC's TD1129 ASCII-character version of surface marine data. Of the 10 duplicate pairs extracted from TD1129, nine were flagged as suspect because of location problems. The remaining pair of duplicate reports flagged as correct did not have a valid ship call-sign and, therefore, the QC track-check was not performed.

We can state with some confidence that the assertion in (3) is correct. Using the binomial probability distribution,it can be determined that, based on our random sample, there is less than a 2% chance that fewer than 60% of the total population of the authors' duplicates would not be caused by location problems and flagged as suspect. The probability is as low as 1 in 500 that 50% or fewer of the duplicates have not been flagged as suspect. Actually, there is a 70% chance that 90% or more of the authors' duplicates can be attributed to location problems and therefore flagged.

The CMR product used by the authors was designed with a compact format which contains only the most frequently used variables. During the creation of CMR, a selection of flags was made to eliminate erroneous data. However, the flag field which indicated suspect position was not used (*i.e.*, the suspect records are in CMR and unflagged). Moreover, CMR was used as input for the calculation of 2-degree latitude by 2-degree longitude monthly summaries (another COADS product).

It is important to note that usually at least one record of each CMR duplicate pair found by the authors was from a telecom source deck. This type of data has been known to be less reliable as compared to logbook data. In particular, the Monterey source (1966-73, as described in *Release 1)* is now thought to be unreliable and should not be used except with extreme caution (NCDC, 1986; 1987).

Based upon the above results, it is recommended that researchers, such as the authors, who require an extremely high degree of duplicate-free data, use either NCDC's TD1129 character product or the packed binary Long Marine Reports (LMR; see Supplement F in *Release 1).* Both of these COADS products contain QC flags which can be interrogated to eliminate erroneous data and some duplicates. Another approach would be to eliminate all telecom data from studies. However, for most surface marine applications, the telecom duplication rate is negligible. Revisions are planned for COADS products in the future (Woodruff *et al.*, 1987), including probable removal of the Monterey source from the 2-degree monthly summaries, but are not warranted at this time for such a minor change.

Additional sources of potential duplicates

The authors discussed one particular aspect of duplicate elimination: a ship's location. However, to eliminate effectively most duplicates, three additional main checks are required between two ship reports. These are the date, time, and individual weather elements.

The DUPELIM procedure developed for COADS was designed to allow less stringent computer checks in those four areas. In addition, when a systematic error was found within a certain data-set, a special allowance was incorporated in DUPELIM.

Supplement K in *Release 1* describes many systematic errors which were circumvented by DUPELIM. However, some problem still remain, and occur mainly near the boundary tolerance of the four main checks in DUPELIM. One problem, similar in magnitude to that of the authors', involved the date of individual observations. In certain data-sets, ship reports were found to be exact matches except that the days differed by one. Allowances were made in DUPELIM, but only within a month. When days crossed into the next month (or year), no attempt was made to eliminate the duplicate. Several areas were tested and revealed a duplication rate of 0.4%. This rate was judged to be insignificant in regard to the expense required for corrections. Similar problems of smaller magnitude exist with location, time, and the weather elements. All are described in detail in *Release 1.*

Conclusion

The authors found a duplication rate in the COADS CMR of less than 0.6%. The duplicates occurred primarily because of differences in whole degrees latitude or longitude of a ship's location. Although a problem was discovered in the authors' processing scheme, most duplicates occur in CMR and are confined primarily to telecom sources. However, when a random sample of these telecom duplicates was selected and compared to NCDC's RD1129 product, most had already been flagged as suspect due to location problems.

It is therefore, recommended that researchers who need an extremely high degree of duplicate-free data use either the TD1129 or LMR products or eliminate all telecom (at the very least, Monterey deck 555) from processing. However, for most surface marine applications, the telecom duplicates found by the authors are negligible. *Release 1* describes additional sources of potential duplicates of similar magnitude.

References

Lander, M.A. and M.L. Morrissey, 1987: Unexpected duplicate ship reports in the comprehensive Ocean-Atmosphere Data Set (COADS). *(Trop. Ocean-Atmos. Newsletter),* No. 38, 13-14.

National Climatic Data Center, 1986: *Marine Data Users Reference, 1970-current,* NCDC, Asheville, NC.

National Climatic Data Center, 1987: *Marine Data Users Reference, 1854-*

1969. NCDC, Asheville, NC.

Slutz, R.J., S.J. Lubker, J.D. Hiscox, S.D. Woodruff, R.L. Jenne, D.H. Joseph, P.M. Steurer, and J.D. Elms, 1985: *Comprehensive Ocean-Atmosphere Data Set; Release 1*. NOAA Environmental Research Laboratories, Climatic Research Program, Boulder, CO, 268 pp. (NTIS PB86-105723).

Woodruff, S.D., R.J. Slutz, R.L. Jenne, and P.M. Steurer, 1987: A comprehensive ocean-atmosphere data set. (Accepted for publication in *Bull. Amer. Meteor. Soc.)*.

Peter M. Steurer
National Climatic Data Center
Asheville, North Carolina